CHALMERS
UNIVERSITY OF TECHNOLOGY

Artificial Intelligence, 3D Documentation, and Rock Art - Approaching and Reflecting on the Automation of Identification and Classification of Rock

Journal of Archaeological Method and Theory
https://doi.org/10.1007/s10816-021-09518-6

Artificial Intelligence, 3D Documentation, and Rock Art—Approaching and Reflecting on the Automation of Identification and Classification of Rock Art Images

Christian Horn, et al. *[full author details at the end of the article]*

Accepted: 28 February 2021/Published online: 12 March 2021

Abstract

Rock art carvings, which are best described as petroglyphs, were produced by removing parts of the rock surface to create a negative relief. This tradition was particularly strong during the Nordic Bronze Age (1700–550 BC) in southern Scandinavia with over 20,000 boats and thousands of humans, animals, wagons, *etc*. This vivid and highly engaging material provides quantitative data of high potential to understand Bronze Age social structures and ideologies. The ability to provide the technically best possible documentation and to automate identification and classification of images would help to take full advantage of the research potential of petroglyphs in southern Scandinavia and elsewhere. We, therefore, attempted to train a model that locates and classifies image objects using faster region-based convolutional neural network (Faster-RCNN) based on data produced by a novel method to improve visualizing the content of 3D documentations. A newly created layer of 3D rock art documentation provides the best data currently available and has reduced inscribed bias compared to older methods. Several models were trained based on input images annotated with bounding boxes produced with different parameters to find the best solution. The data included 4305 individual images in 408 scans of rock art sites. To enhance the models and enrich the training data, we used data augmentation and transfer learning. The successful models perform exceptionally well on boats and circles, as well as with human figures and wheels. This work was an interdisciplinary undertaking which led to important reflections about archaeology, digital humanities, and artificial intelligence. The reflections and the success represented by the trained models open novel avenues for future research on rock art.

Keywords Rock art · Machine learning · Faster R-CNN · 3D documentation · Visualization · Digital humanities

Background and Research Question

During the Scandinavian Bronze Age (1700–500BC), percussive force was used to make images on exposed bedrock. While some rock paintings exist, the rock art in Scandinavia was primarily produced by removing parts of the rock surface to create a

negative relief. Such images are called petroglyphs, engravings, or carvings. Use of the term carving is prevalent despite it technically being wrong since the images were likely made using percussion rather than carving. The tradition of creating pictures on rocks existed in Sweden, Norway, Denmark, and North-Germany. In addition to bedrock outcrops, some boulders and blocks were also used (Fig. 1). The images consist of over 20,000 boats, over 6000 humans, thousands of animals, as well as wagons, wheels, ploughs, and smaller objects including swords, axes, earrings, and shields.

Rock art all over the world has traditionally been documented using a wide array of methods, including tracing, oblique light photography, and rubbings (Nordbladh, 1981). All these methods capture rock art in visually appealing manners. However, the shortcomings of these methods are that they introduce varying degrees of human bias directly into the documentation in a manner that is rarely comprehensible or visible in its full extent to later users. Further, these methods are reductive and discard the third

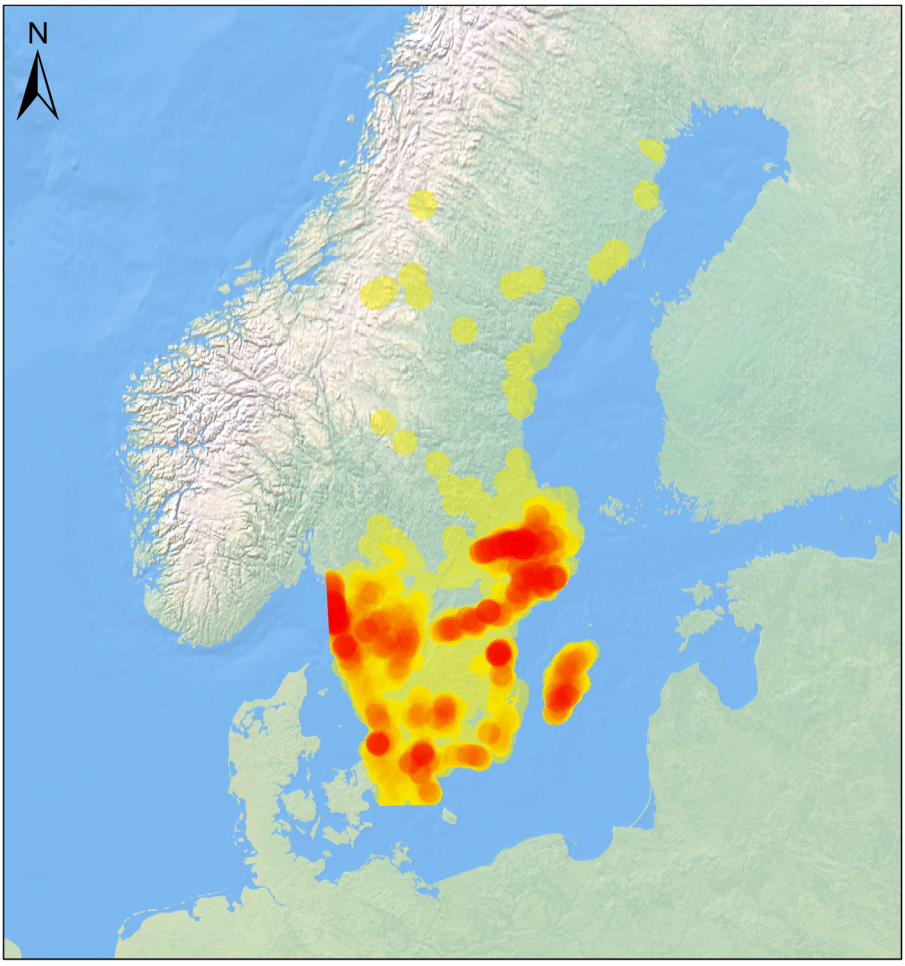

Fig. 1 Distribution of rock art in Sweden (data from the FMIS database of the National Heritage Board for Sweden)

dimension, *i.e.* the depth of the carvings, in the documentation step. This means that they do not contain any reliable, measured depth information. However, recording the third dimension is not only a preservation issue but, as we explain later, also adds opportunities for visualization, post-processing, and research.

In the past decades, the development of more user-friendly and affordable techniques to create 3D models has led to an increasingly stronger inclusion of 3D methods into the panoply of rock art documentation. Methods like laser scanning and Structure from Motion (SfM) record the entire surface without the human operator making any decisions other than selecting an area to document. This means that, within technical limitations, everything in the area is recorded. Even more importantly, the aforementioned methods document rock art in all three dimensions. This is a significant development that has made 3D models the new standard documentation method for rock art (Horn et al., 2018; Mudge et al., 2012). With the recording of three-dimensional models, the amount of data collected increases, which poses new challenges, but also provides opportunities to develop new ways to visualize and analyse the data.

One opportunity was to develop an approach to use the recorded 3D data to automatically detect motifs on rock art panels using artificial intelligence (AI). A method like this could help researchers, heritage institutions, and archives concerned with rock art, such as the Swedish Rock Art Research Archives (SHFA), to automate registration processes. This would make the identification of rock art motifs more consistent and faster and could provide the basic architecture for the development of statistical tools for research, for example, the estimation of the distribution of motifs across chronological phases. It would also make it easier to group all objects from one category together and compare them to estimate stylistic variability. Beyond the technical, a question emerged that intrigued our archaeological and humanist curiosity: How well does AI deal with the complexities of expression of human creativity like rock art, which is often difficult to untangle even for life-long experts? Questions pertaining to the relationship of human and machine-led recognition will not only be of crucial importance to the archaeological field but could also guide the future of the digital humanities in the broader sense.

In the following, we will present a method to automatically identify rock art motifs using 3D data and reflect upon the results. The method is based on new visualization techniques developed explicitly for this purpose, deep learning algorithms, and high-quality 3D documentation. This development work is a collaboration between the SHFA, the County Administrative Board of Västra Götaland, the Centre for Digital Humanities, Chalmers Technical University, and the University of Gothenburg.

Previous Work and Research Question

Projects concerned with the detection of archaeological features in remote sensing data such as DTM and LiDAR data have successfully used deep learning approaches to identify charcoal kilns, burial mounds, Celtic fields (Trier et al., 2015; Trier et al., 2018; Trier et al., 2019), *etc*. There are significant differences in the application of computer vision and AI to rock art, however. A barrow viewed from above, for example, will mostly be a simple shape, *i.e.* round, oval, or elongated. In rock art, the variability of

shapes even for class-like boats is relatively large, despite there being some shared features. For example, a boat may have a keel line or could just be formed from one line; prows may turn outward or inwards and can be symmetrical or asymmetrical; they could have anthropomorphic features like hands or none at all, *etc.* (Fig. 2). Thus, the combination of AI and rock art represents a unique set of challenges and possibilities.

Using AI to analyse rock art is a newly emerging field in both data science and archaeology. In the course of the ERC funded 3D-Pitoti project, several methods have been proposed and tested on a relatively limited set of rock art images (Poier et al., 2016, 2017; Seidl, 2016; Zeppelzauer et al., 2015; Zeppelzauer et al., 2016; Zeppelzauer & Seidl, 2015). 2D and 3D documentations were used to create an automatic segmentation algorithm. In computer vision, image segmentation is used to divide an image into several segments to identify boundaries between objects within the image. In the 3D-Pitoti project, these segments comprised the carved rock and the natural surface of the rock. The project established a library of known features by letting experts annotate the images to enrich their data and evaluate the performance of the algorithm. The project also experimented with the use of their petroglyphs in convolutional neural networks (CNN). Furthermore, the research results demonstrated that geometric information, *i.e.* depth information, performed better in the models than colour information, *i.e.* 2D images. However, to improve the handling of the 3D data, the 3D-Pitoti project also proposed to use of depth maps as a reduced form of 3D data.

Based on this previous work, two major questions could be further investigated: Firstly, is it possible to develop an AI approach using faster and easier annotation which does not necessarily require experts but is still capable of identifying rock art?

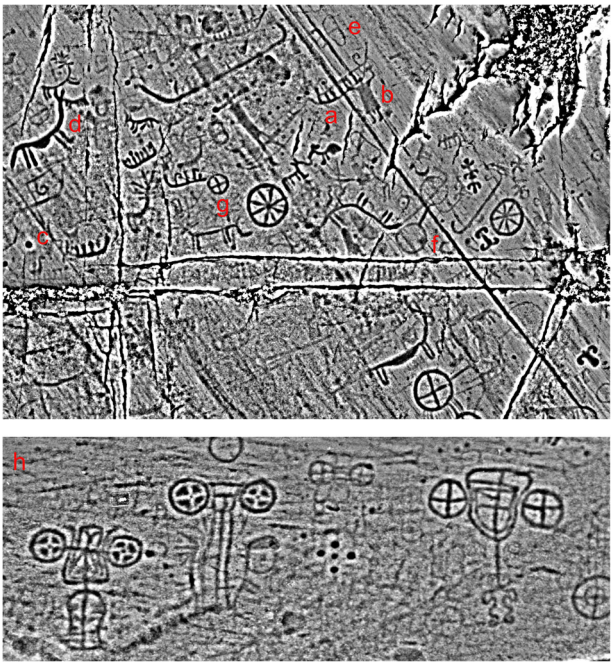

Fig. 2 Examples of the range of motif in rock art: boats (a), humans (b), cupmarks (c), animals (d), footsoles (e), circular features (f), wheels (g), and examples of wagons (h)

Secondly, how can traditional archaeology and more broadly humanistic theorizing reflect on the results and contribute to the future development of such applications? We wanted to develop an approach by using visualizations derived from 3D recordings of rock art and use a larger dataset to train a CNN-based algorithm for object detection. Object detection is a method in computer vision that deals with locating objects in an image and assigns them a class label to identify the object. The problem in our project is that the object needs to be identified and located correctly regardless of scale, position, and orientation. Before we describe our approach, we will address some issues and limitations encountered in our data. Finally, we reflect on the results with a humanistic (archaeological) lens to see what this new approach could contribute and whether it may be able to help understand shortcomings of the method.

Data and Method

The HandySCAN 700™ can provide scan resolutions of up to 0.2 mm and outputs data as a 3D mesh through VXelements. This provides highly detailed scans which record surfaces and reduce human bias inscribed into the documentation. However, as no method is perfect, some shortcomings need to be kept in mind to arrive at the best solution. These scans include natural noise such as exfoliation and erosion patches which leads to the partial loss of petroglyphs and makes the images more difficult to recognize. Such damage on the natural rock also leads to deep cracks that cannot be scanned, which sometimes result in areas that lack, or have sparse, scan data. Similar problems occur at the boundaries of the scan. Partial scans mean that motifs may not be fully recorded. One of the problems with working in an outside environment is that some of the factors that negatively impact scan quality are difficult to control. This includes the reflective nature of grains in the granite, which are recorded by the scanner as null data. Direct sunlight on the panel and/or moisture worsens this problem. The target points necessary for the scan are also recorded as null data. Holes in the scan are more likely to occur on higher resolutions because the data is less interpolated. For this reason, most of the data is close to 1.0 mm in resolution, although 0.2 mm also exists. The HandySCAN 700™ does not record texture information. Therefore, the raw data needs to be rendered, and light needs to be applied to create shadows and highlights to be able to inspect the surface visually.

The data in this project includes 408 laser scans, most of which were scanned rock art located in Bohuslän. Scans from Östergötland, Scania, and Uppland were also used. Of this data, two-dimensional visualizations (see below) of these scans were annotated with bounding boxes. If the sites are named, the inventory number of the Swedish National Heritage Board (RAÄ) is mentioned as well. With this number, the site is searchable on the website Fornsök, and older documentation can be viewed in the online portal of the SHFA.

Visualization

Despite the advantages of the 3D documentation in recognizing and analysing rock art, preserving its third dimension, and the relative independence of viewing and lighting angles, it can be challenging to view and interpret. The 3D model is, after all, just a

representation of a surface, *i.e.* coordinates in a 3D space. To make these more visible, textures can be added, several illuminations can be applied, and advanced rendering techniques such as ambient occlusion can be used. However, the human eye can only visualize two dimensions in the retinal image and requires multiple combined cues, *i.e.* the movement of light, the surface, and the viewer, to perceive 3D shapes (Bertin, 1983; Green, 1998; Welchman et al., 2005). Annotating the 3D model, therefore, carries the risk of overlooking features and makes it harder to control results. Consequently, it was crucial to develop an approach that highlighted all the features in the 3D model independently of movement and lighting. Additionally, 3D data requires high-level computer power. Therefore, the data was projected to two dimensions which provided the following advantages over unstructured 3D data:

- Created the opportunity to use computationally more efficient image analysis methods to avoid out-of-memory issues.
- Allowed us to use the many well-established image recognition methods for 2D data.
- Annotation of the data for supervised rock art recognition tasks using bounding boxes was faster.
- The results of the annotation were easier to control.

The meshes from the scanner have vertices that are uniformly distributed, so it is possible to easily extract a 3D point cloud which we used to create a second point cloud by sampling the mesh. We decided to develop the pipeline of the 3D-Pitoti project by creating visualizations from the 3D model using depth maps (Zeppelzauer et al., 2016). Compared to the 3D-Pitoti data, the Scandinavian data had a stronger global curvature which, in effect, hid features in the depth map (Fig. 3a). We previously developed an easy-to-use visualization approach using methods developed for landscape archaeology (See also Horn et al., 2019). This process creates images that visualize and highlight the content of the panels in great detail showing motifs which are otherwise hard to recognize. However, the parameters of the visualizations are hard to control, and better control of these variables was desired. Therefore, we programmed a new tool ("ratopoviz" = rock art topographic visualization) to automate the creation of visualizations for 3D rock art data. "Ratopoviz" generates depth maps, normal maps, topographic maps, enhanced topographic maps, and blended maps both in colour and greyscale (Fig. 3a–e).

Point Sampling

The first step towards generating images from a 3D mesh was to extract the data in a point cloud that represented the surface of the panel. 3D meshes are convenient to render the data or to apply algorithms where connectivity between data points are essential. For calculating a depth map, it was easier to work with point clouds. Thus, we sampled points from the surface of the mesh. Since the mesh was quite dense and the vertices are evenly distributed through the whole mesh, a simple extraction method was applied by using the vertices in the mesh as points in the point cloud.

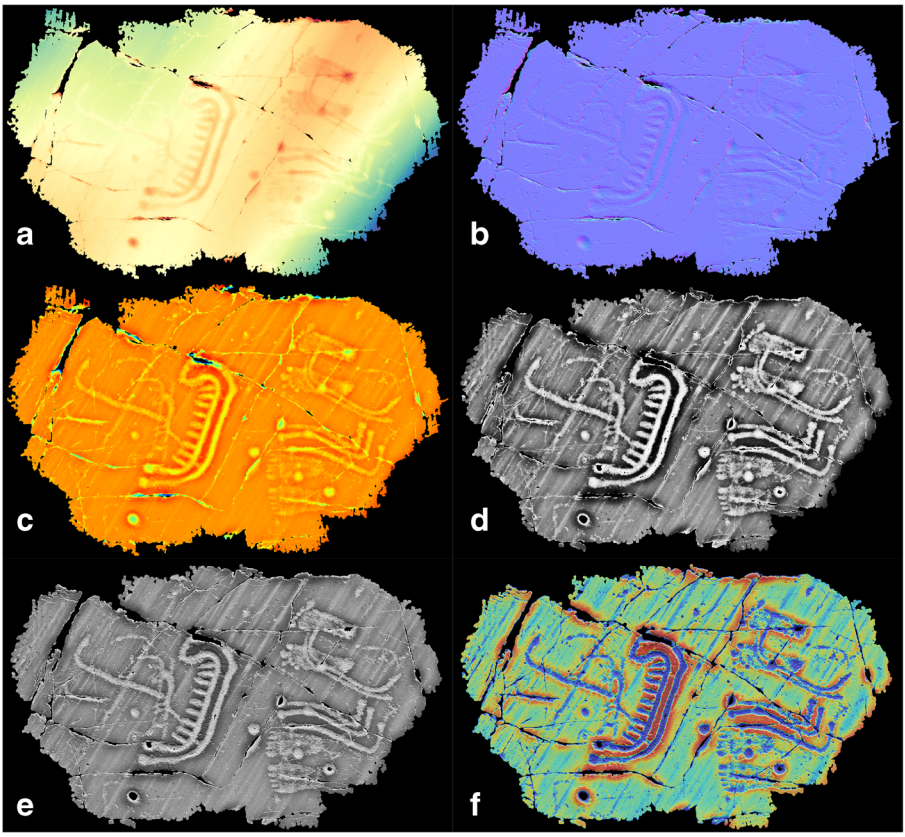

Fig. 3 Depth map (**a**), normal map (**b**), topographic map (**c**), enhanced topographic map (**d**), blended map greyscale (**e**), and blended map colour (**f**) generated using "ratopoviz"

Outlier Removal

To further refine the point cloud data, outliers, and points in sparse areas of the panel were removed. These were, for example, points floating above the surface caused by dust that was recorded during the scanning process, points along edges of the scanned area, and fractures in the rock. To do this, we applied different types of noise detection methods to identify such points and remove them from our point cloud. This was achieved through density-based spatial clustering (DBSCAN), which is a clustering algorithm that groups points which are closely packed together. Since most parts of the panels are composed of regions with dense points, it can be used to detect areas with sparser points. After the clustering process, the algorithm detected smaller clusters with low amount of points which could later be identified as outliers. This method was applied twice, on a local scale of 30×30-cm tiles and on a global scale for the whole panel. Statistical outlier removal detected points that were further away from their neighbourhood than the average for the entire point cloud. These sparse point clusters were then removed.

Principal Component Analysis

To use the point cloud to generate images, it was first necessary to setup a plane to project it onto. This was performed using principal component analysis (PCA). PCA was used to calculate an orthogonal transformation to a new coordinate system so that the first principal component accounted for as much variance in the data as possible. In simple terms, this was the longest vector that could span the panel in the 3D coordinate space. The first two principal components were used to span the plane, and the third was used to store the signed distances from the plane to the original data points. The new coordinate system thus kept the same dimensions as the original.

Generating Images

To create the images, it was necessary to make a projection from 3D to 2D and go from a set of unstructured points to a set of structured points, *i.e.* a regular grid of pixels. This was performed by using the first two principal components as the *x*- and *y*-axis in the image and use the signed distances from the third component to generate pixel values. Representing data with a smaller number of dimensions would cause a loss of information. The data used in this project was convenient to represent in 2D since the panels were one-sided and close to flat in their global surface curvature. For this reason, there is also no recognizable distortion.

An image resolution was chosen based on the resolution of the 3D data. Most of the panels had a resolution close to 1.0 mm in 3D space, and the resulting pixel size in the generated image was set to the mesh resolution divided by two. This means that the number of pixels along an axis in the generated image would be approximately twice the amount of points, *e.g.* a panel with a width of 1 m will be of 2000 pixels. The new pixel values were set by linear interpolation of the signed distances from the third principal component, *i.e.* depth values. Additionally, normal vectors from the 3D data were also interpolated and mapped to RGB channels to generate normal maps. A normal in 3D space is a vector which is perpendicular to the surface. The resulting normal maps store information about the directions of each pixel.

The resulting depth maps had some areas that only contained pixel values interpolated from very distant data points. They were not part of the actual scanned panel. This was because a convex hull was created around the panel during the interpolation phase. To remove the value of these pixels, we computed the distance of each pixel to the three nearest neighbours in the point cloud. If that distance was over a given threshold, set by the average and standard deviation of all distances in the point cloud, then the value of the pixel was set to NaN (not a number).

Removing Global Curvature

All panels had a more or less pronounced global curvature, *i.e.* the overall shape of the bedrock. If the curvature was weak, the visualization of the local variations was largely undisturbed. However, if the surfaces were larger and/or the global curvature stronger, it disturbed and hid the actual patterns that we wanted to reveal in the panel, *i.e.* the rock art. A blurred, smoothed version of the image was created which did not contain the local variation but represented the global curvature. Extracting this from the original

image resulted in a version that only contained the local variations (Fig. 3c). This variation contained natural striations (ice-lines), damage like natural cracks, and the rock art motifs. There was a trade-off between the amount of information and the convenience of working with it. However, it was feasible since we could keep a relatively high resolution based on real-life depth data.

The smoothed version was generated through Gaussian blurring, using the same method as Zeppelzauer and Seidl (30). A parameter needed to be set related to the standard deviation of the distribution is used in the Gaussian function. This parameter controlled the amount of smoothing in the image and needed to be set individually for each image if one wanted to keep as much detail as possible for the resulting image. Since we dealt with panels of roughly the same resolution, it was enough to set this parameter once, depending on the scale according to which the image will be smoothed. In our case, it was on an object level since we wanted to highlight single motifs in the panels.

Image Processing

The output of the process explained above displays the local variation in depth through the panels based on absolute depth values. This variation, however, can be minimal. Therefore, when performing visual inspections of the panels, it might be challenging to identify all the rock art because it is disturbed, for example, by extreme values and natural noise. In this step, we tried to improve the images to highlight the content further.

First, extreme values were removed at a lower and upper threshold. The thresholds were selected based on the 25th and 75th percentile of the data. However, the removal was adjusted with a value multiplied at the interquartile range. This value was selected through testing and visual inspection of the result on a smaller set of panels of different sizes and types. Smaller isolated areas in the image were also removed. To further minimize the effect of extreme values, a logarithmic scaling was applied. The scaling was followed by a histogram equalization on a local scale through the CLAHE method (contrast-limited adaptive histogram equalization). This was to distribute the intensities on all possible values and thus increase the contrast in the image (Fig. 3d).

The Data and Annotation

The 408 original 3D meshes were projected down to 2D images. The images were available in greyscale and with a red-yellow-green colour map. One reason to represent the data in 2D format was for visualization purposes, another was that it was less time consuming to annotate the data on an instance level with bounding boxes, and a third was that many well-established methods exist for processing 2D images when it comes to object detection. The 3D-Pitoti project used shape annotation (Seidl, 2016); however, this is a time-consuming technique and could result in issues given the large amounts of data required for machine learning.

In line with this, we chose a bounding box approach which mirrors the region proposals made by the object detection, which were also output as bounding boxes. This method is recognized as easy to use and a time-saving process (Dai et al., 2015). Bounding boxes denoted regions in which a motif is located. Afterwards, the motif was

assigned a class label. The classification was kept relatively simple to not break up the data too much. The classification had eleven initial labels: boat, human, animal, wagon, cupmark, footsole, circle, wheel, rider, script, and other. The annotation was carried out using the tool "labellmg" which saved the bounding box data in the PASCAL VOC format. The annotations were transferable between all visualizations of the same scan.

The annotation was performed on a basic level with the main classifications that exist in Swedish rock art (*e.g.* Fig. 2). Such classifications of the annotation labels are inherently difficult as the images are often ambiguous (Cabak Rédei et al., 2020) and multiple potential identifications could be made (see Fig. 6). To account for this problem, the classification was based on the most readable images first, gradually annotating more difficult panels. Two additional rock art researchers checked the label classification to reduce labeller bias. All 408 scans were annotated. Partial figures that were located at the borders of the scan or petroglyphs heavily affected by erosion were marked as difficult. This was relevant for around 10% of the objects. These are not yet considered for the object detection task, which means 4286 objects were identified in all. Classes with very few samples, *i.e.* script (2) and rider (17), were included in the group other (300). This meant that these objects were not well represented in the data so it might be difficult to identify them. Although plentiful, cupmarks were also excluded as they could be overrepresented as a simple object that can easily be recognized by simpler means and because they could cause confusion when they are a part of more complex petroglyphs, for example, human figures frequently use cupmarks as representations of their heads (Horn, 2016). In all, five of the 10 original categories were selected as main classes to test object detection: boat (612), human (808), animal (294), circle (91), and wheel (98) (see Fig. 4). There are many animals with different shapes depicted on the Scandinavian rock panels, *i.e.* cattle, snakes, fish, birds, and deer. Therefore, only four-legged animals were included in this classification.

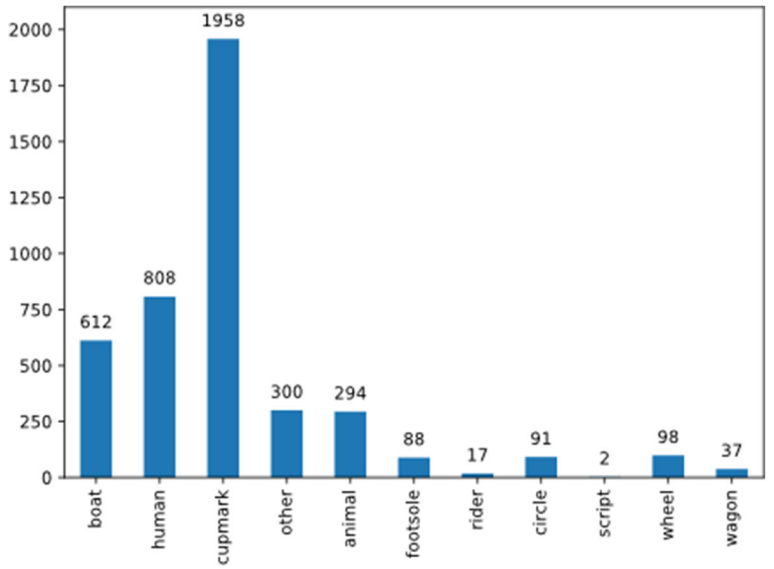

Fig. 4 Table showing the member size for each object class

The data included rock art sites from different areas in Sweden, including Scania, Östergötland, and Uppland. Most scans were taken on rock art sites in Bohuslän by the SHFA and the County Administrative Board of Västra Götaland. Each site may have several panels. Such panels could have been scanned several times in different versions or they could be scanned in overlapping sections. The latter was done when the scan of the panel would have become too data intensive to allow proper calculation by the scanning software, and there was an expectation of subsequent problems in exporting and handling such large files. To avoid having the same objects in the training and the testing data, the data were first grouped by panel, and the resulting groups were then split into train (~80%), test (~10%), and validation (~10%) sets. The validation set was used during the training to save the model with the lowest validation loss.

One future extension would be to further increase the test and validation set when more data is available or implement cross-validation to conduct a comprehensive estimation of the predictive performance of the model. However, it is not possible to use cross-validation while testing different hyperparameters. Additionally, training already takes a long time, and time constraints precluded the use of such an approach for this paper.

Faster Region-Based Convolutional Neural Network

Faster region-based convolutional neural network (Faster R-CNN) (Ren et al., 2016) is an object detection network that takes an image as an input and produces bounding boxes with class labels. In our setup, the inputs are scaled versions of the image types extracted from the 3D data with corresponding bounding boxes and class labels from the annotation. The images were scaled due to memory constraints and had a shortest edge of 600 pixels. Faster R-CNN works as a two-stage cascade which means that after the first step, in which a Region Proposal Network (RPN) proposes regions of the image that most likely contain objects, the region proposals are further adjusted, and a final classification of these object regions is calculated.

The architecture of Faster R-CNN can be divided into three sub-networks. The first network (feature extraction network), illustrated as red in Fig. 5, processes the input image and extracts features into feature maps. All feature maps stacked together have a depth of 1024, and each cell corresponds to a receptive field of approximately 25–50% of the input image. In our approach, this network was constructed using the first four building blocks of the ResNet-50 (He et al., 2016) architecture. ResNet-50 achieves high accuracy because of its depth, but it is still simple to train due to its shortcut connections. The building blocks of the network contain multiple convolutional layers followed by batch normalization and non-linear activation functions. The shortcut connections skip these blocks making it possible for gradient information to bypass these blocks during backpropagation. This type of residual network resolves the problem of vanishing gradients that can affect performance when training deeper networks. Another advantage of ResNet is that it exists in different pre-trained versions using the ImageNet dataset (Deng et al., 2009; Krizhevsky et al., 2012), which is convenient for transfer learning (see the "Transfer Learning" section).

The resulting feature map is fed into the RPN (green in Fig. 5) to propose regions of the image that most likely contain objects. The RPN consists of three layers. First a convolutional layer is used to map each window of 3×3 cells in the feature map to a

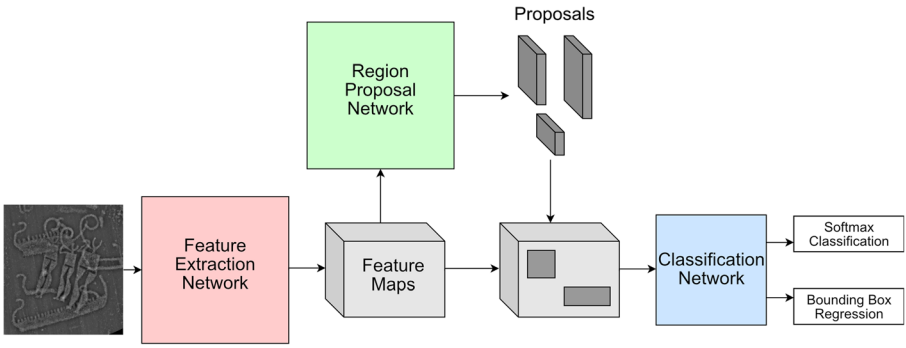

Fig. 5 Structure of the Faster R-CNN workflow used by the project

512-dimensional feature vector. These feature vectors are then used as the input into two sibling convolutional layers, one for a box regression with a linear activation function and the other for a box classification layer with a sigmoid activation function. The output is a set of rectangular bounding boxes with a related probability for the box including an object.

The classification network (blue in Fig. 5) uses the proposed regions from the RPN and the feature maps to perform region of interest (RoI) pooling. Because the outputs from the RPN can be of varying shapes and sizes, it is necessary to make them a fixed size before the final dense layers. RoI takes the section of the feature maps that corresponds to a region proposal and scales it to a fixed size of 14×14 cells. The final layers of the classifier include the following:

- A few convolutional layers with shortcut connections
- Average pooling to summarize the average presence of features in patches of 7×7 cells
- Two outputs layers, one with linear activation for further refinement of the bounding box and the other with softmax activation to get the probability distribution for the class labels

For each input image, the output is a list of bounding boxes with corresponding probabilities for the class labels used in training (see "The Data and Annotation" section), together with an additional label corresponding to the background in the image when no object is present.

Faster R-CNN was mainly selected for the task of rock art detection because it has been used with success in other archaeological projects. Verschoof-van der Vaart and Lambers (2019) used it to detect prehistoric barrows and Celtic fields in LiDAR data. Trier and his colleagues (2015, 2018, 2019) used the method on DTM data for the automatic detection of grave mounds, charcoal kilns, etc. and for the semi-automatic mapping of archaeological landscapes. Despite these successes, there were no guarantees that the method would work on rock art as the other projects used it for remote sensing and implemented machine learning for the detection of large-scale features in this data. In contrast, image features associated with rock art can be variable and complex, providing an entirely different set of features.

Data Augmentation

Deep learning models like Faster R-CNN require a high volume of data to perform well. These powerful models can find complex non-linear relationships in the data and learn high-level features that are useful in several object recognition tasks. However, the model contains plenty of internal parameters. This means that such models can easily overfit when the dataset is small, *i.e.* the error on the data used for training is small, but the generalization is poor, making the error on the test data large. This results in poor predictive performance.

The risk for overfitting can be remedied by expanding the dataset through the augmentation of the existing data to increase the size and diversity of the dataset. For example, in the re-projection of the 3D data to 2D, the images can rotate, and with that, the objects also rotate. It is therefore important to train the object detector on several different rotations. Another basic augmentation technique is tiling the image into multiple sub-images with a specified overlap between the tiles. One reason for this is that the original images often have high resolutions. Measured from the shorter edge, smaller scans may have ca. 1000 pixels, while larger scans may have up to over 10000 pixels. Processing them fully is memory intensive and led to out-of-memory errors. However, down-sampling directly risks severe data loss. Therefore, the larger images were divided into tiles that could be down sampled with the risk of only minor data loss. Tiles are created in with a maximum size of 2000 pixels and an overlap of 400 pixels. Because of the overlap, multiple outputs can be generated for the same object when performing predictions on the whole image. In these cases, an average of all overlapping bounding boxes for the same class label is computed.

Other augmentation techniques were applied randomly involving the following geometric and colour operations (Fig. 6):

- Horizontal and vertical flipping
- 90° rotation
- Arbitrary rotation by a maximum of 3°
- Shearing
- Brightness adjustment
- Contrast adjustment
- Adding salt and pepper noise
- Adding Gaussian noise
- Adding Poisson noise

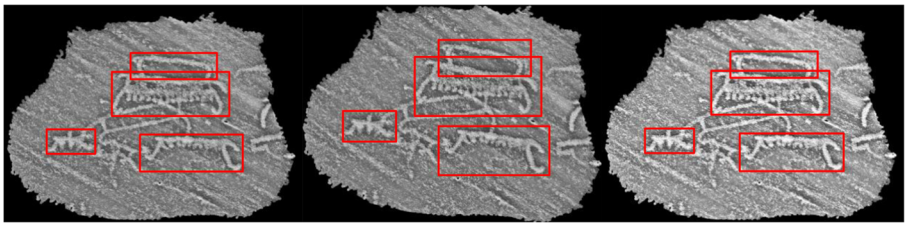

Fig. 6 Augmentations of an original visualization (left): slightly rotated (centre) and contrast adjustment (right)

Transfer Learning

Transfer learning was also used to improve generalization and avoid overfitting. Transfer learning has been widely applied in machine learning tasks instead of, and in combination with, data augmentation to reduce overfitting, particularly where there are limited training datasets available and networks have been fine-tuned, for example, in medical imaging to detect Alzheimer's disease (Oh et al., 2019). In a first step, the network is pre-trained on a large-scale dataset, such as ImageNet data, to learn general features that can later be used in a second step with fine-tuning on a specific target task. Even though the target task uses a different data type, transfer learning can provide generalization of features like edge detection that would be similar in large and very small datasets. Transfer learning was performed in 3 steps:

1. The layers that are shared between the RPN and the classification network within the architecture (Faster R-CNN) are initialized with weights from a pre-trained ResNet-50 on the ImageNet dataset. The final layers of the RPN and the classification network that are not shared are randomly initialized.
2. The shared layers are then frozen, and the whole network is fine-tuned during 100 epochs.
3. A few of the first layers in the shared layers remain frozen, while the others become unfrozen, and the whole network is further trained for at least 300 epochs.

Prediction

Since each training sample is a tile of an image (see "Data Augmentation" section), the prediction also needs to be performed on tiles. Since the tiles overlap, one object can be found in several predictions. This was solved by using a version of non-maximum suppression. Instead of taking the bounding box with highest predicted confidence, the bounding boxes that overlap by an IoU[1] of more than 0.2 were averaged into one single final bounding box and probability.

Results

First, we compare the effects of data augmentation and transfer learning. The model was trained for 100+300 epochs, and the input data were greyscale images whose intensity distributions were equalized. The values were expressed as an average precision for each object class and as a mean average precision (mAP) for all object classes. The usage of an augmented and pre-trained network increases the generalization capacity, and thus, the model has a higher predictive performance (see Table 1 part a).

Furthermore, different image input types made from the same scan were compared (Fig. 3c–f). The first type had only the global curvature removed. The second was the same used for the comparison above, with the global curvature removed and

[1] Intersection over Union: area of overlap divided by area of union for two bounding boxes

Table 1 Performance comparison (mAP) for unaltered models, augmented data, and additional transfer learning (a); performance comparison (mAP) for the different image inputs (b); performance comparison (mAP) for multiple image inputs (c)

a	No Augmentation No Transfer Learning	With Augmentation No Transfer Learning	With Augmentation With Transfer Learning
Animal	0.0	4.8	8.8
Boat	4.7	42.0	58.0
Circle	0.0	7.1	31.4
Human	0.0	10.2	25.3
Wheel	0.0	0.0	21.6
mAP	1.0	12.9	29.0

b	Image Type 1 Curvature Removal	Image Type 2 Transformation	Image Type 3 + Blending with Normal Map
Animal	5.6	8.8	7.0
Boat	40.1	58.0	51.3
Circle	57.1	31.4	22.2
Human	27.7	25.3	21.5
Wheel	21.6	21.6	15.9
mAP	30.4	29.0	23.6

c	Multi Type 2 (Type 1 + Type 2)	Multi Type 3 (Type 1 + Type 2 + Type 3)
Animal	7.1	10.0
Boat	64.1	61.3
Circle	40.0	33.3
Human	28.9	24.0
Wheel	22.2	18.8
mAP	32.5	29.5

further equalization applied to highlight all local deviations. The third type had been additionally enhanced by blending the output with the normal map. All three image types were in greyscale. While many may find the type 3 images most appealing and the best to read because they are conveying a 3D "feel", they did not perform well. Looking at the mAP, simple curvature removal seems to have the best predictive performance. However, for the classification of animals and boats, type 2 images provided the best generalization capacity (see Table 1b).

This means that the performance of the input types differed depending on the class label. Therefore, it was decided to use multi-type training to see if the performance could be further improved. Table 1 part c only represents the results of the multi-type training in which such improvements were shown. Combining type 1 and 2 images gave the best results, recognizing almost two-thirds of all boats and a third of the human figures. Combining all three image types yielded the best performance for recognizing animals. However, the average precision for animals was exceptionally low overall.

Further tests were performed by increasing the resolution of the input images, training with other sets of hyperparameters, and training on the full images. Still, it did not lead to any consistent improvements from the models above.

Discussion

From the table and diagram (Table 1 part c; Fig. 7), it is possible to see that data augmentation and transfer learning produced a model that avoided strong, direct overfitting. However, there was still quite a large difference between the training and the validation loss (Fig. 8). This meant that some overfitting still occurred even though the loss on the validation data did not become drastically worse. This was probably due to the large variation of objects in the data, a relatively small dataset, and the use of a complex model with a lot of parameters. The main labels used for comparison of the different models were human and boat. Except for cupmarks, these were the two most common types of objects in the data. It is difficult to draw any conclusions from the other object classes since they are somewhat underrepresented in the data.

Another issue to notice for the precision-recall curves is that they never reach the point where we have a recall value of 1.0 (Fig. 7). This implies that we are never able to capture all objects in the test data with our predictions. One thing to try would be to adjust the IoU threshold, which was set to 0.7, and evaluate if a lower threshold also means higher recall values. However, this was outside the scope of this paper.

Upon visual inspection of the results, it became clear that it was harder for the model to predict larger objects with one single bounding box. Since we averaged overlapping bounding boxes of the same class, the reason for this could be that bounding boxes on a larger object did not overlap to a sufficient degree. The label with the highest probability was plotted, which could cause the bounding box prediction with the highest likelihood to be somewhat misaligned with the motif (Fig. 9).

Different image types as input perform better on different object classes, and a combination performs better than each one individually. As was previously mentioned, type 2 images perform better to recognize boats, while type 1 images work better for humans (Table 1 part b). Images of type 2 have equally distributed pixel values which highlight noisy parts of the scans. One reason could be that human figures are often comparatively small and consist of fewer lines than boats making them harder to separate from the noisier parts of the scan. If that noise is enhanced, then it becomes even harder to recognize them. Boats, on the other hand, are often larger objects and seem to benefit from the equalization. A future model could be trained on both types of images and use the benefits from both datasets. However, it could also be used as an augmentation technique.

Overall, the object class “animal” has a seemingly poor performance. A reason for this could be that even among the four-legged animals there is considerable variation, for example, horns, tails, genitalia, length of legs, and length of necks. This together with the rather limited number of training examples could have caused the issue. This could also indicate another reason why the model performed less well in recognizing humans than boats. Generally, in rock art research, human or other anthropomorphic figures are identified as a combination of the bodily features and whichever objects are associated with the figure.

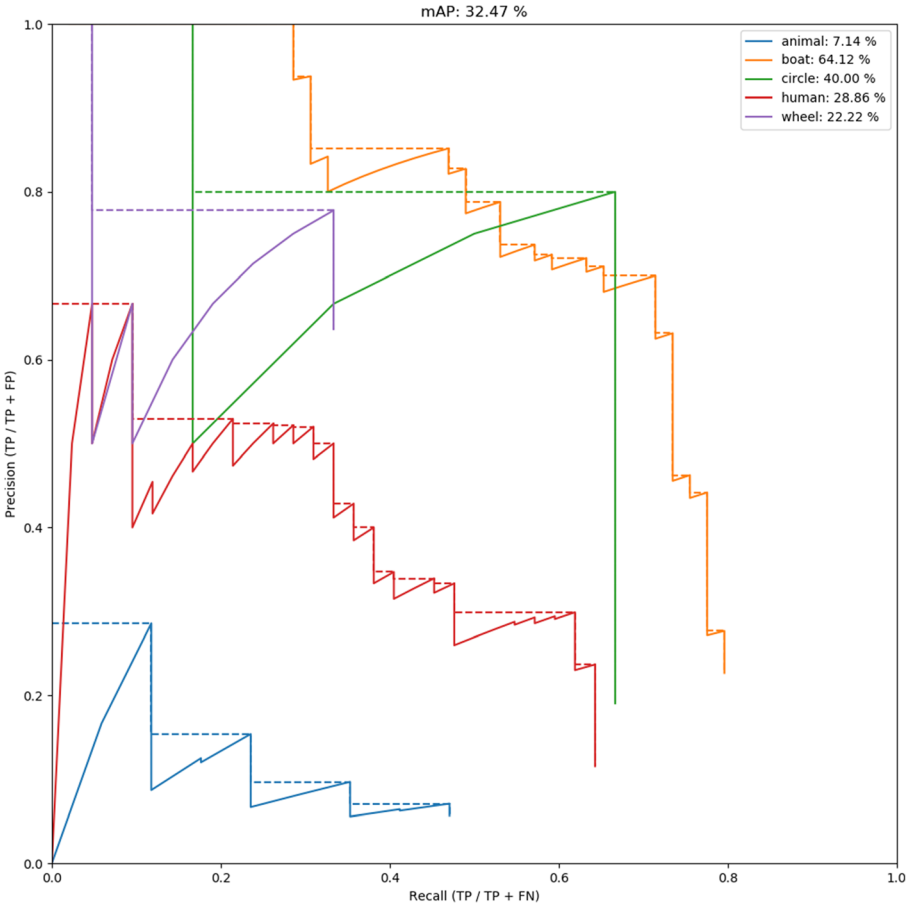

Fig. 7 Precision-recall curves for the multi-type 2 model

Reflection

Reflecting on these results from an archaeological perspective is to recognize how difficult a material like rock art is. Petroglyphs are very complex, and they are the outcome of equally complex imaginations, ideologies, traditions, and belief systems made by individuals informed by social institutions. In some cases, these expressions of human creativity are further complicated through changes made by later carvers (Bertilsson, 2015; Horn & Potter, 2018; Ling & Bertilsson, 2017). Furthermore, the shapes of rock art motifs are very fluid in two respects. One aspect is that the shape of some carvings closely resembles the shape of other carvings; animals can, for example, closely resemble boats (Fig. 10a). The second aspect of the fluidity of rock art is that humans, boats, and/or animals sometimes intersect in a way that means that body parts are replaced by features of the boat or animal (Fig. 10b) (Horn, 2018). In this, it is interesting that, at least for now, the CNN has trouble handling this kind of ambiguity and, in that sense, with human creativity.

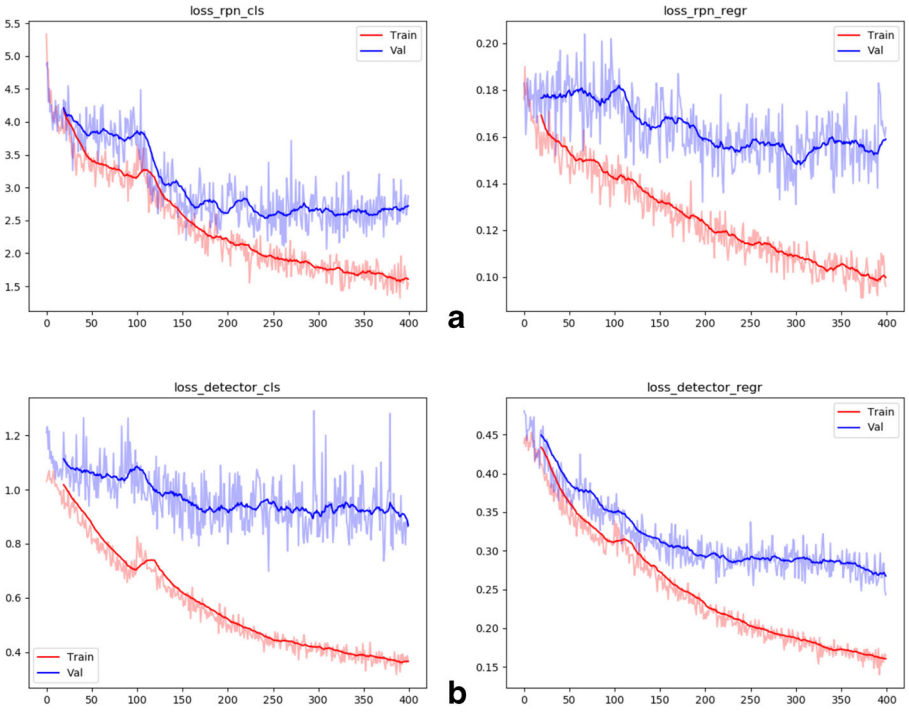

Fig. 8 The training (red) and validation (blue) loss for the RPN network (**a**) and the classification network (**b**) for the multi-type 2 model

However, when comparing the visualizations, with the annotations, and then the computer-made predictions, some extraordinary observations were made. On a scene (Tanum 248:1) with three lur-blowers[2], the algorithm marked a small area above the three humans as a boat with 96% confidence. This is indeed a boat which had been left out of the annotation because it was weakly carved and close to the edge of the scan and was therefore not clear that all parts of the boat were in the scan (Fig. 11). That means the algorithm predicted something correctly that was not previously annotated, which is in itself a good result. Upon further inspection, it was discovered that on a panel with two warriors and a boat, the CNN predicted another boat (Fig. 9). This is another carved area that was not annotated: within the glacial line, there is part of a circle or wheel which was not documented in its entirety by the scanning and was therefore left out of the annotation.

This meant the CNN was able to recognize motifs it was not trained on, which offers the potential that it may predict motifs that were not recognized when the rock art was recorded or later analysed. An example of this was discovered on another panel from Kville (149:2). The panel had a large boat with wheels above it, and also a pair of wavy lines often interpreted as snakes (Fig. 12). Previous documentations of this panel had left the area at the bottom next to the boat empty. When annotating it again, it was recognized that the area was somewhat eroded (Fig. 12). The discovery of new images was not the aim of annotation; therefore, no in-depth visual inspection was undertaken.

[2] Lurs are Bronze Age wind instruments that have been made from bronze.

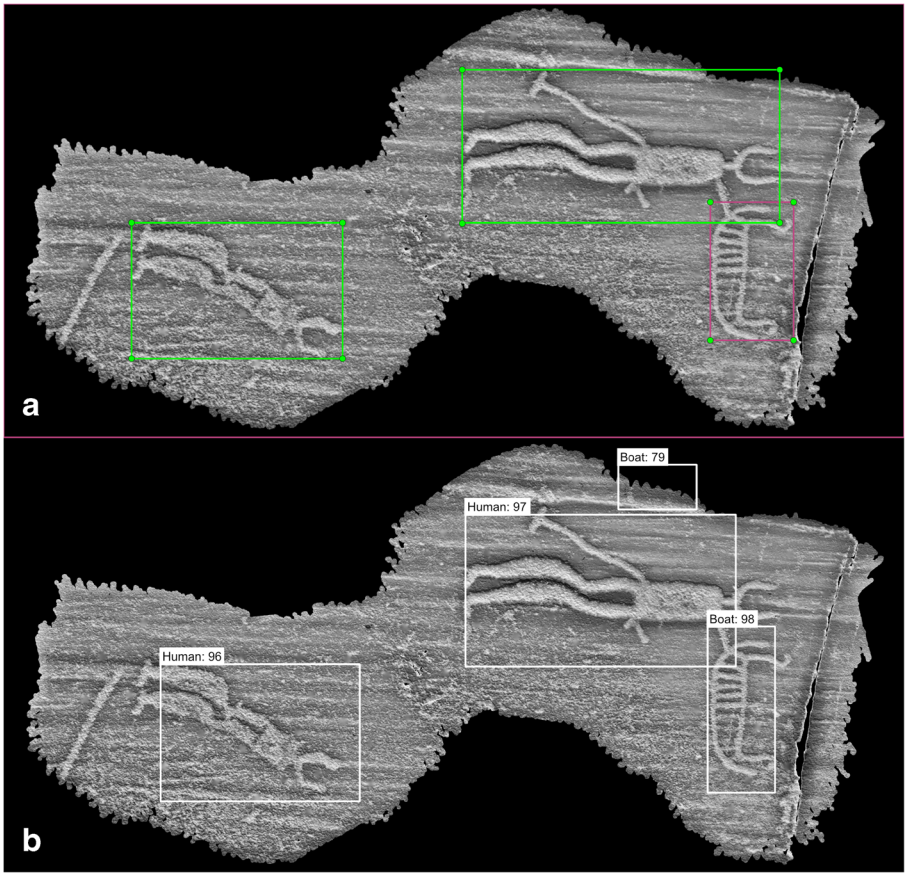

Fig. 9 Kville 12:1 annotation (**a**; boxes: green = human, pink = boat) compared to predictions (**b**)

In that area, the CNN predicted the presence of a human figure (96%) and an animal (89%). Upon closer inspection of the 3D file and various visualizations, it was confirmed that there are legs and at least an upper body (Fig. 13). This seems to overlay another line which could be the baseline for the carving representing the phallus and sword sheath line often observed on warrior figures (Horn, 2018). The left leg goes over into what the algorithm identifies as an animal (Fig. 12). If it is assumed that the figure under the human has the same orientation, and then it should instead be interpreted as a boat.

This highlights that the CNN struggled with the aforementioned fluidity of forms of Scandinavian rock art. Some animals, if turned 180°, look very similar to boats (Fig. 9a). The same is true for some humans when turned 90°. The algorithm cannot know which is the "correct" upper side, because it is not trained for this. In addition, in rock art, it is not always clear for all motifs which the "right" viewing direction should be (Janik, 2014). The ambiguity of rock art is highlighted by another prediction the algorithm made on Kville 149:2. There is a smaller boat lower on the scanned area which the CNN correctly recognized as a boat. For the right prow of the boat, the algorithm

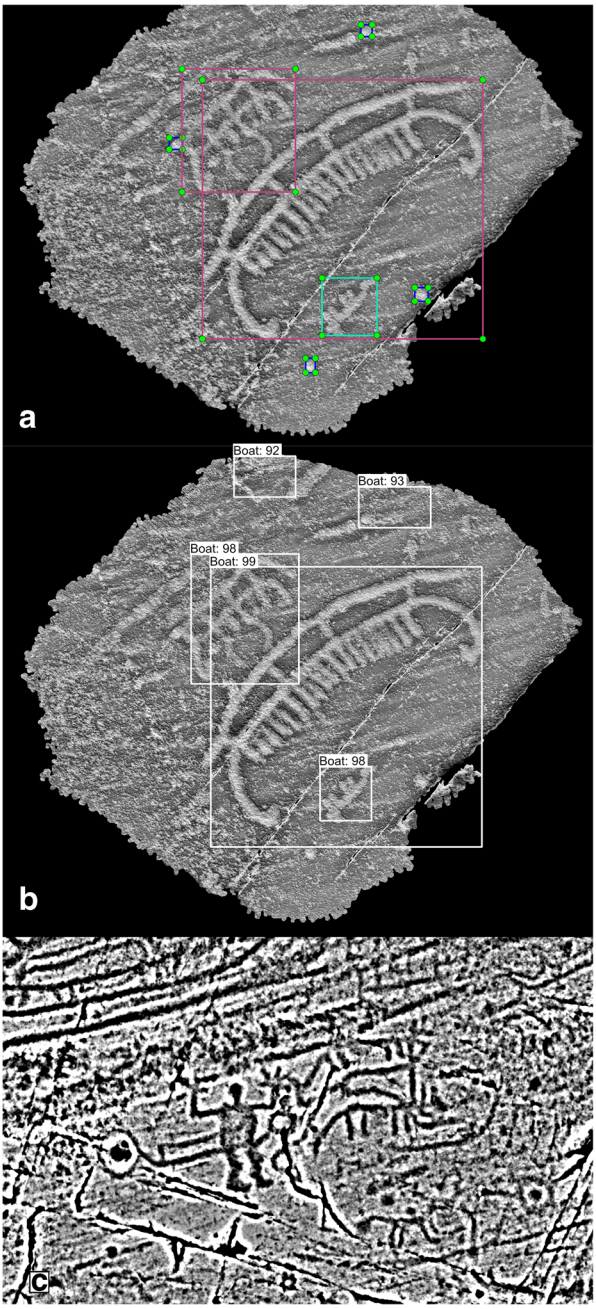

Fig. 10 Small animal on Tanum 25:1 recognizes as a boat, because upside down it looks like a boat (**a** annotation; boxes: green = human, pink = boat, blue = cupmark; **b** prediction), human intersecting a boat so that the stem forms the phallus on Bottna 74:1 (**c**)

predicted an animal (Fig. 12). Although it may seem odd, it is not wrong. From about 1500 to 1400 BC, prows were formed with animal heads, mostly horse

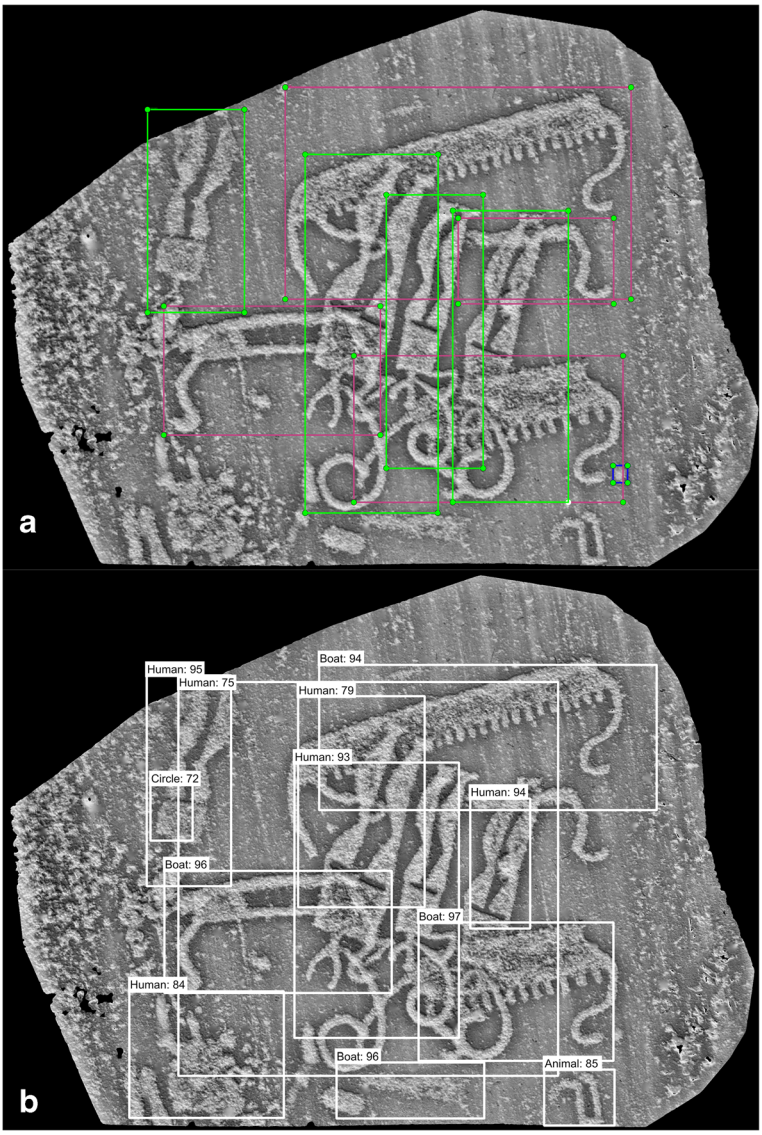

Fig. 11 Tanum 248:1 annotation (**a**) compared to prediction (**b**)

heads. The prow itself and the side of the boat can seem like the front of a horse carving.

The future of artificial intelligence approaches to rock art depends on the possibilities of theorizing about the creativity of making rock art in environments in which we lack direct information on the intentions behind their production and their meanings. In the wider sense, the problems and possibilities of interpreting and creating meaning are considered to be the main limitation for today's artificial intelligence (Steels & Wahle, 2020). The algorithm learns and predicts only what we teach it. Thus, we have to find a scale. Should the boat in Kville be recognized as a boat with a horsehead, as a boat and

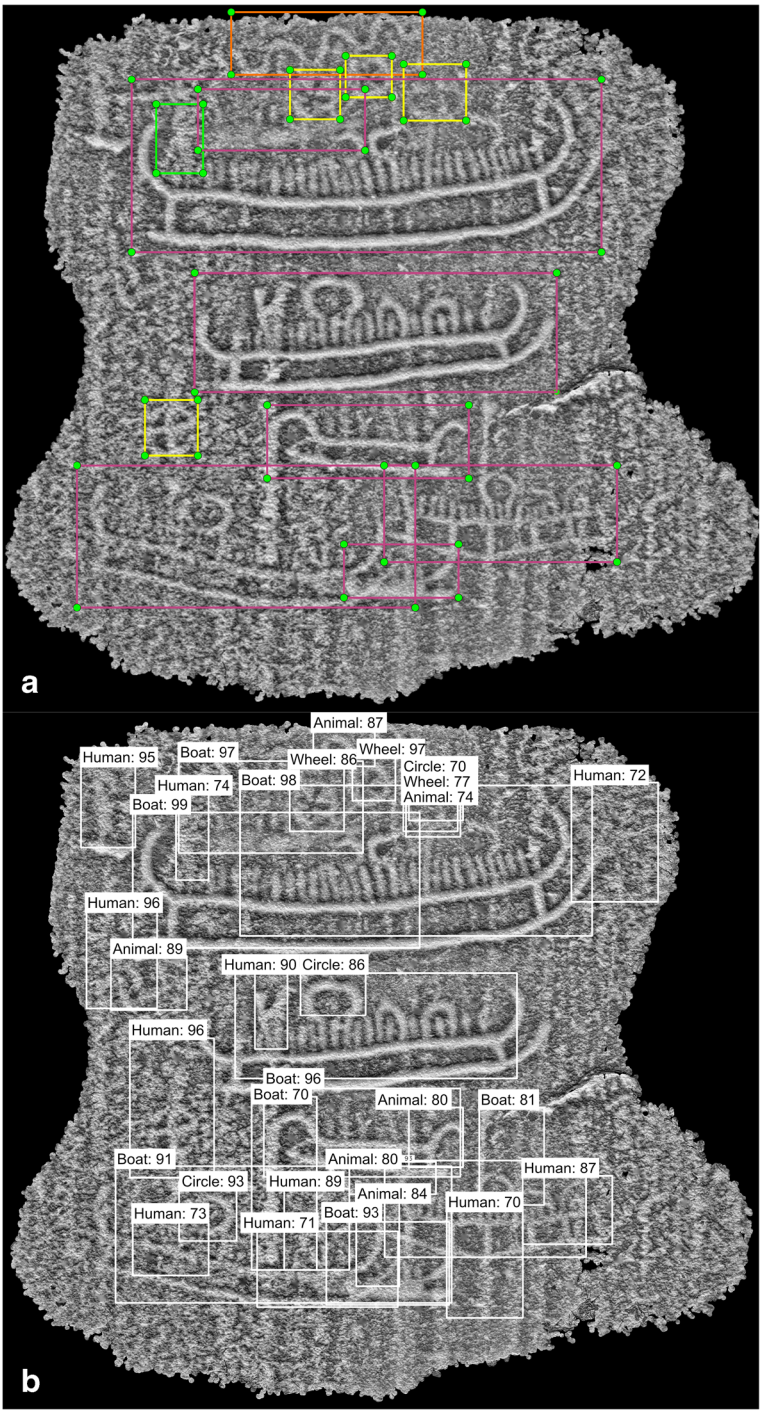

Fig. 12 Kville 149:2 annotation (**a**; boxes: green = human, pink = boat, orange = animal, yellow = wheel) compared to prediction (**b**)

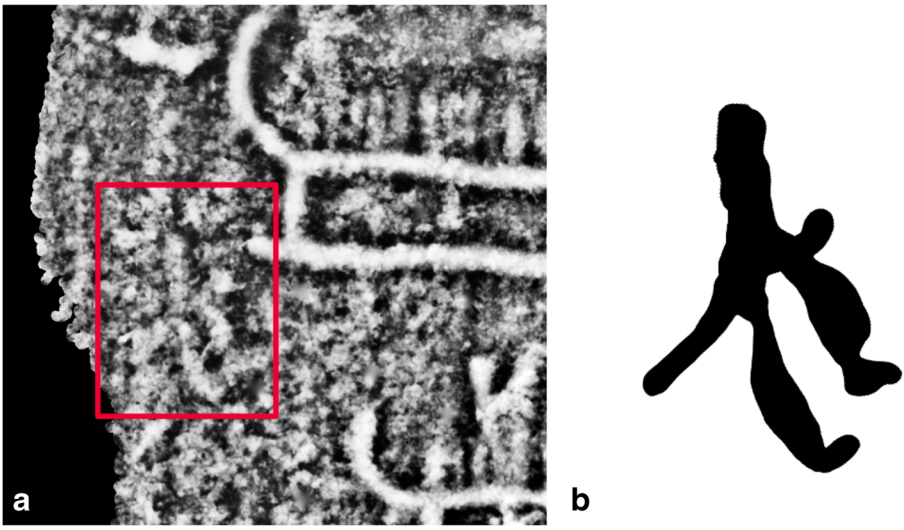

Fig. 13 Area of the potential human on Kville 149:2 marked (**a**); drawn interpretation of the human (**b**)

a horse, or as a hybrid creature distinct from other boats? Depending on the different views taken, the outcome of any AI approach and its interpretation will differ. In turn, this enables us to reflect on the potential limitations of our digital technology to avoid taking its results as answers to our research questions about human beliefs, ideologies, and creativity.

Conclusion

Despite some difficulties and setbacks, this project was successful, and the trained CNN can be used for future endeavours. It can recognize major motif classes, and it can be expected to be more accurate in detecting and classifying other objects when more training data becomes available. For an automated keywording system, the CNN is already usable to speed up the workflow since a keyword needs to be given only once per panel even if a motif occurs multiple times. Since motifs like boats, circular features, and humans are often present multiple times, the CNN will recognize it in one or two out of three cases. That means some human control is necessary, especially for animals and other features that occur on fewer occasions.

Reflecting on the results in general and on a case by case basis provides interesting challenges that digital archaeologists will have to discuss when applying artificial intelligence to the products of human creativity, especially when any direct information on meanings and intentions are missing.

Acknowledgements We are thankful to Henrik Zedig (County Administrative Board of Västra Götaland) for making laser scans produced by him and his team available to us and the non-scientific staff of the Swedish Rock Art Research Archives for their tireless work. We would like to extend our thanks to the two anonymous peer reviewers for their excellent feedback on this paper. Victor Wåhlstrand Skärström gave insightful advice to the machine learning component of this project for which we are grateful. All errors remain of course our own. Furthermore, we are indebted to the Bank of Sweden Tercentenary Foundation for supporting our work.

Availability of Data and Material All 3D models and visualizations are available under CC BY-NC-ND 4.0 license from the SHFA upon request. They will also continuously be uploaded to https://sketchfab.com/SHFA-3D.

Code Availability The code and executable for ratopoviz is available at: https://github.com/Swedish-Rock-Art-Research-Archives/rock-art-ratopoviz. The code for the rock art detection model can be downloaded at: https://github.com/Swedish-Rock-Art-Research-Archives/rock-art-radnet.

Funding Open access funding provided by University of Gothenburg. This work was conducted in the project "Rock art in three dimensions" funded by the Bank of Sweden Tercentenary Foundation under grant no. IN18-0557:1.

Declarations

Competing Interests The authors declare no competing interests.

References

Bertilsson, U. (2015). Examples of application of modern digital techniques and methods: Structure for motion (SfM) and multi-view stereo (MvS) for three-dimensional documentation of rock carvings in Tanum creating new opportunities for interpretation and dating. In F. Troletti (Ed.), *Prospects for prehistoric Rock Art research*: *XXVI Valcamonica Symposium* (pp. 57–62). Capo di Ponte: Centro camuno di studi preistorici

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps*. Eurospan.

Cabak Rédei, A., Skoglund, P., & Persson, T. (2020). Seeing different motifs in one picture: Identifying ambiguous figures in South Scandinavian Bronze Age rock art. *Cogent Arts & Humanities, 1802804*. https://doi.org/10.1080/23311983.2020.1802804.

Dai, J., He, K., & Sun, J. (2015). BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision, 2015*, 1635–1643

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In IEEE Staff (Ed.), *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), Miami, FL, 6/20/2009 - 6/25/2009* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848.

Green, M. (1998). Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Human Factors Science*, 8

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2016, 770–778

Horn, C. (2016). Cupmarks. *Adoranten, 2015*, 29–43.

Horn, C. (2018). Fast like a war canoe: Pragmamorphism in Scandinavian rock art. In A. Dolfini, R. J. Crellin, C. Horn, & M. Uckelmann (Eds.), *Prehistoric Warfare and Violence: Quantitative and Qualitative Approaches* (pp. 109–127). Cham: Springer.

Horn, C., & Potter, R. (2018). Transforming the rocks – Time and rock art in Bohuslän, Sweden: Time and rock art in Bohuslän, Sweden. *European Journal of Archaeology, 21*(3), 361–384. https://doi.org/10.1017/eaa.2017.38.

Horn, C., Ling, J., Bertilsson, U., & Potter, R. (2018). By all means necessary: 2.5D and 3D recording of surfaces in the study of southern Scandinavian rock art. *Open Archaeology, 4*(1), 81–96. https://doi.org/10.1515/opar-2018-0005.

Horn, C., Pitman, D., & Potter, R. (2019). An evaluation of the visualisation and interpretive potential of applying GIS data processing techniques to 3D rock art data. *Journal of Archaeological Science: Reports, 27*, 101971. https://doi.org/10.1016/j.jasrep.2019.101971.

Janik, L. (2014). Seeing visual narrative: New methodologies in the study of prehistoric visual depictions. *Archaeological Dialogues, 21*(1), 103–126. https://doi.org/10.1017/S1380203814000129.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Ling, J., & Bertilsson, U. (2017). Biography of the Fossum Panel. *Adoranten, 2016*, 58–72.

Mudge, M., Schroer, C., Noble, T., Matthews, N., Rusinkiewicz, S., & Toler-Franklin, C. (2012). Robust and scientifically reliable rock art documentation from digital photographs. In J. McDonald & P. M. Veth (Eds.), *A companion to rock art* (pp. 644–659). Wiley-Blackwell.

Nordbladh, J. (1981). Knowledge and information in Swedish petroglyph documentation. In C.-A. Moberg (Ed.), *Similar finds? Similar interpretations?: Glastonbury - Gothenburg - Gotland ; nine essays* (pp. G1–G79). Univiserty of Gothenburg.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., & Oh, I.-S. (2019). Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports, 9*(1), 18150. https://doi.org/10.1038/s41598-019-54548-6.

Poier, G., Seidl, M., Zeppelzauer, M., Reinbacher, C., Schaich, M., Bellandi, G., Marretta, A., Bischof, H. (2016). PetroSurf3D: A high-resolution 3D dataset of rock art for surface segmentation. *arXiv preprint*

Poier, G., Seidl, M., Zeppelzauer, M., Reinbacher, C., Schaich, M., Bellandi, G., Marretta, A., & Bischof, H. (2017). The 3D-Pitoti dataset. In ACM (Ed.), *the 15th International Workshop, Florence, Italy, 19.06.2017* (pp. 1–7, ICPS). The Association for Computing Machinery. https://doi.org/10.1145/3095713.3095719.

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv, 1506.01497*, 1–14

Seidl, M. (2016). *Computational analysis of petroglyphs*. PhD. Technische Universität Wien, Vienna

Steels, L., Wahle, B. (2020). *Perceiving the focal point of a painting with AI: Case studies on works of Luc Tuymans*. Institute for Systems and Technologies of Information, Control and Communication

Trier, Ø. D., Zortea, M., & Tonning, C. (2015). Automatic detection of mound structures in airborne laser scanning data. *Journal of Archaeological Science: Reports, 2*, 69–79. https://doi.org/10.1016/j.jasrep.2015.01.005.

Trier, Ø. D., Salberg, A.-B., & Pilø, L. H. (2018). Semi-automatic mapping of charcoal kilns from airborne laser scanning data using Deep Learning. In M. Matsumoto & E. Uleberg (Eds.), *CAA2016: Oceans of data proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology* (pp. 219–231). Archaeopress Publishing Ltd..

Trier, Ø. D., Cowley, D. C., & Waldeland, A. U. (2019). Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeological Prospection, 26*(2), 165–175. https://doi.org/10.1002/arp.1731.

Verschoof-van der Vaart, W. B., & Lambers, K. (2019). Learning to Look at LiDAR: The Use of R-CNN in the automated detection of archaeological objects in LiDAR Data from the Netherlands. *Journal of Computer Applications in Archaeology, 2*(1), 31–40. https://doi.org/10.5334/jcaa.32.

Welchman, A. E., Deubelius, A., Conrad, V., Bülthoff, H. H., & Kourtzi, Z. (2005). 3D shape perception from combined depth cues in human visual cortex. *Nature Neuroscience, 8*(6), 820–827. https://doi.org/10.1038/nn1461.

Zeppelzauer, M., & Seidl, M. (2015). Efficient image-space extraction and representation of 3D surface topography. In *2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015-09-27* (pp. 2845–2849). IEEE. https://doi.org/10.1109/ICIP.2015.7351322.

Zeppelzauer, M., Poier, G., Seidl, M., Reinbacher, C., Breiteneder, C., Bischof, H., & Schulter, S. (2015). Interactive segmentation of rock-art in high-resolution 3D reconstructions. *Digital Heritage – IEEE, 2*, 37–44. https://doi.org/10.1109/DigitalHeritage.2015.7419450.

Zeppelzauer, M., Poier, G., Seidl, M., Reinbacher, C., Schulter, S., Breiteneder, C., & Bischof, H. (2016). Interactive 3D segmentation of rock-art by enhanced depth maps and gradient preserving regularization. *Journal on Computing and Cultural Heritage, 9*(4), 1–30. https://doi.org/10.1145/2950062.

Affiliations

Christian Horn[1] · Oscar Ivarsson[2] · Cecilia Lindhé[3] · Rich Potter[1] · Ashely Green[4] · Johan Ling[5]

✉ Christian Horn
christian.horn@gu.se

[1] Department for Historical Studies and Swedish Rock Art Research Archives, University of Gothenburg, Gothenburg, Sweden

[2] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

[3] Centre for Digital Humanities, University of Gothenburg, Gothenburg, Sweden

[4] Department for Historical Studies and Centre for Digital Humanities, University of Gothenburg, Gothenburg, Sweden

[5] Department for Historical Studies, University of Gothenburg, Gothenburg, Sweden